GESTIÓN ASISTIDA DE DOCUMENTOS EN UNA METODOLOGÍA DE EXPLOTACIÓN DE INFORMACIÓN

E. Fernández[1,2], H. Merlino[1,2], M. Ochoa[1,2], E. Diez[1], P. Britos[1] y R. García-Martínez[1]

[1]Centro de Ingeniería de Software e Ingeniería del Conocimiento. Escuela de Postgrado. Instituto Tecnológico de Buenos Aires
[2]Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires.

rgm@itba.edu.ar

Resumen: El abordaje metodológico de un proyecto de ingeniería del software se apoya en un conjunto de actividades que generan documentación y que exige el uso de herramientas software que permitan realizar una gestión de documentos completa, ordenada y con capacidad de evolución. El presente trabajo describe las características de una herramienta software que permita gestionar la documentación de proyectos de explotación de datos basada en la metodología CRISP-DM, permitiendo a múltiples usuarios trabajar dentro de un mismo proyecto y sirviendo de guía para el desarrollo de los distintos documentos que componen el proyecto.

Palabras claves: CRISP-DM, Asistente Metodológico, Metodología de Explotación de Datos, Documentación del proceso de desarrollo.

Workshop: Ingeniería de Software y Bases de Datos (WISBD).

1- Introducción

Para llevar a cabo un proyecto de Ingeniería del software, cualquiera sea su enfoque, es fundamental contar con una metodología de desarrollo [Somerville, 1998; Pressman, 2002], la cual se apoya en un conjunto actividades que generan documentación [Diez *et al.*, 2003]. Esta forma de trabajo, si bien no asegura el éxito del proyecto, permite administrar de forma eficiente los recursos asignados al mismo. La mayoría de las metodologías de desarrollo actuales se basan en ciclos de vida de desarrollo iterativos. Tal es el caso de RUP [IBM, 2005], Métrica Versión III [2004] o CRISP-DM [Chapman, *et al.*, 1999] entre otras. Para poder aplicar esta forma de trabajo se debe contar con alguna herramienta software que permita realizar una gestión de documentos completa (que permita resguardar todos los documentos del proyecto), ordenada (que permita asociar cada uno de los documentos con la actividad que los generó) y con capacidad de evolución (que permita el versionado de los distintos documentos que generan). Por otra parte dentro de cada una de las fases de desarrollo es normal encontrar un conjunto de subfases o documentos opcionales, los cuales dependiendo de las características del proyecto que se está desarrollando podrán ser completado o no. Por último, es importante destacar que en la actualidad, salvo raras excepciones, los proyectos de ingeniería del software que se desarrollan en las distintas empresas difícilmente se asignen a una única persona. Por lo cual, el proceso de gestión de documentación que se aplique al proyecto deberá permitir vincular a cada uno de los elementos del proyecto con su responsable. El presente trabajo describe al desarrollo de una herramienta software que permita gestionar la documentación de proyectos de Explotación de Datos (data mining) basada en la metodología

CRISP-DM, permitiendo a múltiples usuarios trabajar dentro de un mismo proyecto y sirviendo de guía para el desarrollo de los distintos documentos que componen el proyecto.

2- Estado de la Cuestión

Hoy día son cada vez mas las empresas y organizaciones que para poder dar respuesta a preguntas como: ¿Qué características particulares tienen mis clientes? o ¿Cómo se comportarán mis clientes ante determinada decisión comercial? o ¿Qué patrones de comportamiento poseen los pacientes de determinada enfermedad?, intentan buscar una solución, a estas preguntas, en los datos almacenados en sus bases de datos históricas. Esto ha hecho que los desarrollos de sistemas destinados a realizar procesos de explotación de datos sean cada ves mas frecuentes y complejos. Para poder construir un sistema de este tipo a gran escala es fundamental contar con una metodología acorde al problema. Cabe destacar la experiencia reflejada por los creadores de la metodología CRISP-DM, en la cual destacan el interés mostrado por la audiencia, durante un workshop de data mining [CRISP, 1997], respecto de cual o que metodología de desarrollo se debería aplicar para llevar a delante un proyecto de estas características. Lo que demuestra un alto estado de conciencia respecto de todas las ventajas que aporta trabajar de forma ordenada y documentada. Fruto de las inquietudes de estos participantes surgió: CRISP-DM, una metodología estándar que ha sido desarrollada para la construcción de proyectos de Explotación de Datos. Creada por un consorcio de compañías, principalmente europeas, y su nombre significa: Cross-Industry Standard Process for Data Mining. Aunque se desarrollo para llevar adelante grandes proyectos, es suficientemente amplia y flexible para aplicarla a proyectos de cualquier tamaño. En la figura 1 se detalla el ciclo de vida de un proyecto de Explotación de Datos.

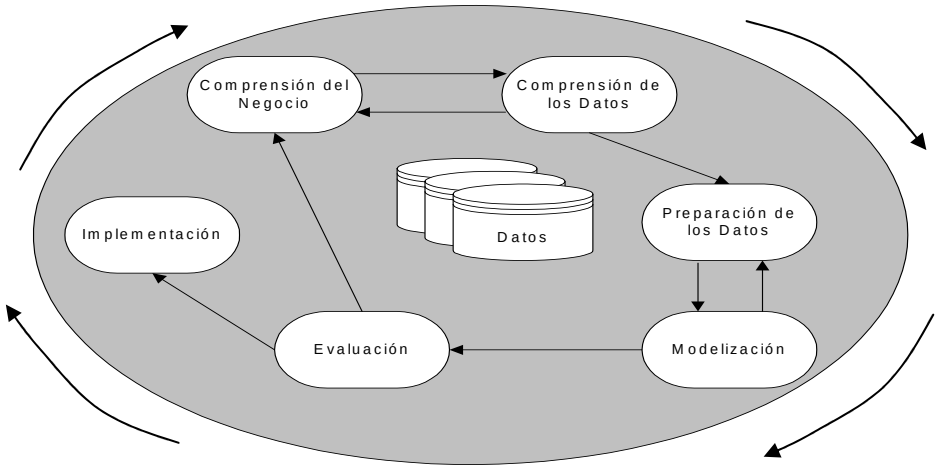

Figura 1. Ciclo de Vida de un Proyecto de Explotación de Datos

El ciclo de vida de un proyecto de Explotación de Datos consiste en seis fases cuya sucesión no es rígida, y se puede mover entre ellas siempre que se requiera. Las flecha indican las dependencia mas importantes y frecuentes entre las fases. El circulo exterior simboliza la naturaleza cíclica de los proyectos de Explotación de Datos. La metodología se presenta en términos de un proceso jerárquico. Consiste en un juego de tareas descriptas en niveles de abstracción (de lo general a lo específico): la fase, la tarea genérica o subfase, la tarea especializada y el caso del proceso.
El contexto de CRISP-DM se maneja entre lo genérico y el nivel especializado, dentro del cual se distinguen cuatro dimensiones diferentes:

- Dominio de la aplicación. Especifica el área en que el proyecto de explotación de datos tiene lugar.
- Tipo de problema. Describe la clase y objetivos del proyecto.
- Aspecto técnico. Cubre procesos específicos de la explotación de datos, describe diferentes desafíos que normalmente ocurren.
- Herramienta técnica especifica que se aplica durante el proceso de explotación de datos.

A continuación, en la tabla 1, se detallan las fases que componen a la metodología CRISP-DM junto con sus subfases y actividades:

Fase 1: Comprensión del negocio	Determinar los objetivos del negocio	▪ Background ▪ Objetivos del negocio ▪ Criterios de éxito del negocio
	Evaluación de la situación	▪ Inventarios de recursos ▪ Requisitos, supuestos y requerimientos ▪ Riesgos y contingencias ▪ Terminología ▪ Costos y beneficios
	Determinar Objetivos del Data Mining	▪ Las metas del Data Minino ▪ Criterios de éxito del Data Mining
	Realizar el Plan del Proyecto	▪ Plan de proyecto ▪ Valoración inicial de herramientas
Fase 2: Entendimiento de los datos	Recolectar los datos Iniciales:	▪ Reporte de recolección de datos iniciales
	Descubrir datos:	▪ Reporte de descripción de los datos
	Exploración de los datos:	▪ Reporte de exploración de datos
	Verificación de calidad de datos	▪ Reporte de calidad de datos
Fase 3: Preparación de los datos	Preparatorios	▪ Dataset ▪ Descripción del dataset
	Seleccionar los datos	▪ Inclusión/ exclusión de datos
	Limpiar los datos	▪ Reporte de calidad de datos limpios
	Estructurar los datos	▪ Derivación de atributos ▪ Generación de registros
	Integrar los datos	▪ Unificación de datos
	Formato de los datos	▪ Reporte de calidad de los datos
Fase 4: Modelo	Seleccionar una técnica de modelado	▪ La técnica modelada ▪ Supuestos del modelo
	Generar el plan de pruebas	▪ Plan de pruebas
	Construir el modelo	▪ Seteo de parámetros ▪ Modelo ▪ Descripción del modelo
	Evaluar el modelo	▪ Evaluar el modelo ▪ Revisación de seteo de parámetros
Fase 5: Evaluación	Evaluar Resultado	▪ Valoración de resultados mineros con respecto al éxito del negocio ▪ Modelos aprobados
	Proceso de revisión	▪ Revisión del proceso
	Determinar Próximos pasos	▪ La técnica modelada ▪ Listar posibles acciones

Fase 6: Implementación	Plan de Implementación	▪ Plan de Implementación
	Plan de monitoreo y mantenimiento:	▪ Plan de monitoreo y mantenimiento
	Informe Final	▪ Informe final ▪ Modelos aprobados
	Revisión del proyecto	▪ Documentación de la experiencia

Tabla 1. detalle de las fases CRISP-DM y sus subfases y actividades

Si bien los creadores de CRISP-DM desarrollaron una metodología amplia y flexible para poder dar soporte a este tipo de proyectos, no se dispone de una herramienta que permita dar soporte a toda la documentación interviniente en un proyecto desarrollado con esta metodología. Por otra parte, actualmente existen herramientas de gestión de documentación para metodologías de proyectos de desarrollo de sistemas convencionales, dentro de las cuales se puede mencionar a Gesmet [2001]. Una herramienta de gestión de documentación de proyectos basados en la metodología Métrica Versión III [Métrica, 2004], la cual permite a múltiples usuarios trabajar remotamente (a través de una red de computadoras) en un mismo proyecto de forma controlada y coordinada. Este sistema, como se muestra en la figura 2, posee una pantalla principal de proyectos en la cual, mediante un árbol de directorio, se muestra cada una de las fases de la metodología, las subfases que componen estas fases, las tareas que se deben realizar dentro de las subfases y los documentos que se generan como resultado de haber realizado la tarea. Es importante destacar que esta herramienta no exige un orden en el llenado de los documentos, ni que se completen todos los documentos definidos en la metodología, esto es debido a que la metodología Métrica Versión III, es una metodología flexible y adaptable, que permite que se obvien algunas tareas y documentos dependiendo de las características del proyecto a tratar en ese momento. Esta característica también es aplicable a CRISP-DM. Otro punto importante a destacar de esta herramienta es la ayuda que provee, no solo sobre el uso mismo de la herramienta, sino también, sobre la metodología que soporta. En la figura 3, se muestra la pantalla principal de ayuda del sistema. Dentro de la cual se muestra la opción de ayuda de Métrica Versión III, desde donde se puede consultar, para cada una de las fases de la metodología, que objetivo tiene la misma, que subfase y actividades la componen, y cual es el propósito y contenido de los documentos a generar.

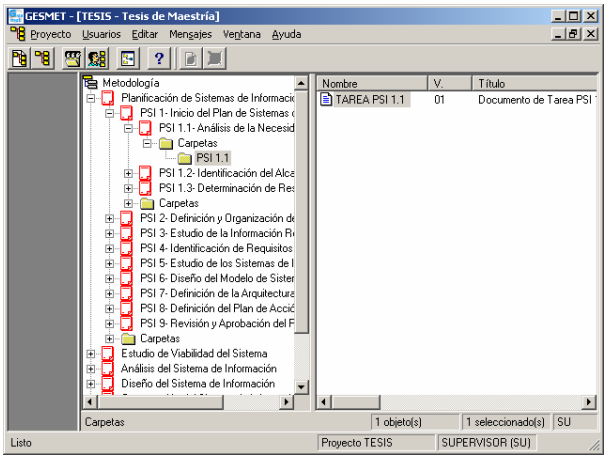

Figura 2. Pantalla principal de proyectos de GESMET

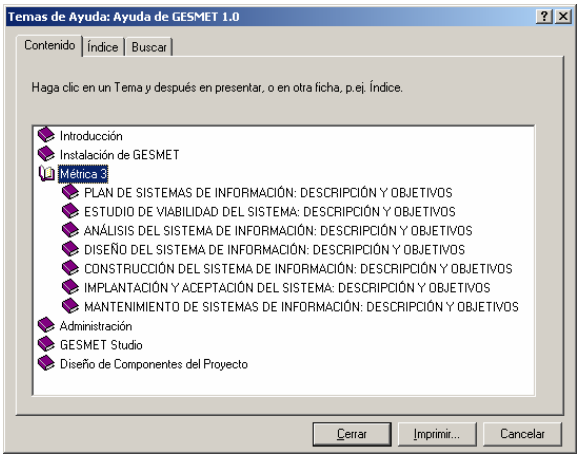

Figura 3: Pantalla principal de ayuda de GESMET

3- Definición del Problema

El presente trabajo está enfocado a solucionar dos problemas importantes en el desarrollo de sistemas de Explotación de Datos basados en la metodología CRISP-DM, estos problemas son:

1. Gestionar toda la documentación del proyecto
2. Brindar ayuda en línea sobre los aspectos mas significativos de la metodología.

Respecto del primero de los puntos, en las secciones de introducción y estado de la cuestión se han dado varios motivos por los que se considera importante esta cuestión. Podemos decir que sin documentación no puede haber una buena planificación; y sin gestión de documentación no puede haber ordenamiento en los documentos y por lo tanto no se puede realizar el seguimiento de la planificación. Si bien la tarea de gestión de documentación, cuando el proyecto es pequeño, puede realizarse manualmente o llevarse en la cabeza de alguno de los integrantes del equipo de desarrollo, cuando el proyecto crece y el grupo de trabajo es asignado a mas de un proyecto estas formas de gestión se tornan caóticas y la forma mas eficiente de solucionarlo es mediante la aplicación de una herramienta software que de soporte a esta función. El segundo punto indicado apunta a obtener una herramienta que no solo permita gestionar los documentos del proyecto, sino que además, brinde información referente a la metodología de desarrollo para que quienes participen en el desarrollo del proyecto puedan sentirse seguros respecto de la información que vuelcan al mismo. Para lo cual es fundamental contar un módulo de ayuda que brinde información sobre cada una de las tareas y documentos que propone la metodología.

4- Solución Propuesta

Se está desarrollando una herramienta software del tipo multiusuario que permita gestionar la documentación de varios proyectos de explotación de datos (de forma paralela) basados en la metodología CRISP-DM. La misma cuenta con:

- Interfaz gráfica de usuario, el sistema cuenta con una interfaz gráfica interactiva que muestra la estructura de los proyectos mediante un menú tipo “árbol de directorio”, donde se pueden apreciar todas las fases de la metodología CRISP-DM. Así mismo para cada fase se debe mostrar, con el mismo criterio, sus “n” subfases a las cuales se puede ingresar mediante la expansión de la Fase. Asimismo, dentro de cada subfase se puede observar los distintos documento que la misma posee, detallando las versiones existentes del mismo e indicando que documentos han sido cumplimentados como datos mas significativos.

- Control de acceso, mediante el ingreso de un nombre de usuario y clave se restringe el acceso de los usuarios al sistema para evitar el ingreso de usuarios no autorizados al mismo.

- Cuatro perfiles diferentes de usuarios (Administrador, Supervisor, Líder de Proyecto y Desarrollador), de esta forma pueden participar del proyecto desarrolladores con diferentes habilidades y responsabilidades, los cuales cuentan con opciones de menú diferente en función del perfil de usuario que tienen asignado.

- Asignación dinámica de usuarios al proyecto, se permite a los usuarios con perfil Supervisor (quienes son los responsables del proyecto para el sistema) asignar y reasigna usuarios de menor nivel jerárquico a las distintas fases y subfases de desarrollo del proyecto. De esta manera el sistema puede registrar quienes son los usuarios autorizados a ingresar a los distintos documentos que posee el proyecto, llevando a demás registro de quienes han ingresado a los mismos.

- Versionado automático de documentos, el sistema permite generar nuevas versiones de documentos, para ello, cuando el usuario lo considera necesario puede indicar que requiere una

nueva versión de un determinado documento. Ante este pedido el sistema genera automáticamente un nuevo documento que contiene toda la información del documento anterior y un nombre similar a este con el número de versión incrementado en uno.

- Una base de datos que centraliza la información de todos los proyectos, la cual permite tener almacenado en un solo lugar la información de todos los proyectos facilitando entre otras tareas la de backup y control de acceso al sistema.

- Un menú de consultas y listados, desde donde los usuarios habilitados pueden obtener información de los distintos proyectos que se están llevando a cabo. Dicha información es restringida en función del perfil de usuario. Así, un usuario con perfil de Supervisor puede obtener datos de los proyectos que el coordina, mientras que un usuario con perfil de Administrador puede ver como evolucionan la totalidad de los proyectos registrados en el sistema, esta última información es útil para los niveles mas altos de decisión de la empresa.

- Capacidades de ampliación a Multiplataforma, si bien la primer versión que se ha desarrollado solo está probada y homologada para correr dentro del entorno Windows, se prevee una versión del sistema en el lenguaje Java para ejecutarse desde distintos entornos Hardware y Software.

- Funciones de ayuda en línea, las cuales abarcan, no solo cuestiones de uso del programa, sino que también, proveen información sobre la metodología CRISP-DM, de forma similar a como Gesmet apoya a los proyectos basados en la metodología Métrica Versión III.

Podemos decir que desde el punto de vista de la gestión del proyecto, esta herramienta permite solucionar una de las tareas mas tediosas del proyecto: el versionado de los documentos y el almacenamiento de la documentación histórica del mismo. Así como también brinda una ayuda en línea que le permite a los usuarios evacuar todo tipo de dudas respecto de la metodología CRISP-DM y el uso de la propia herramienta.

5- Interfaz de la herramienta

A continuación se detallan un conjunto de pantallas y reportes que permiten observar como será la operatoria del sistema propuesto. Es de hacer notar que no se describirán todas las pantallas del sistema (no es la intención mostrar el manual de usuario del sistema), sino que, solo se mostrarán las mas representativas desde el punto de vista de la administración del proyecto. Estas pantallas se vinculan con las funciones de: acceso y creación de un documento y de asignación de usuarios al proyecto.
En la figura 4 se muestra la pantalla de administración de proyectos, desde la cual se puede acceder a cada uno de los documentos del proyecto, consultarlos, editarlos, modificarlos, crear nuevas versiones o dar por finalizado el proyecto. Para ello, la pantalla se compone de dos grillas, en la de la izquierda se despliegan todas las fases que propone la metodología CRISP-DM para un proyecto, y en la de la derecha, en función de la subfase seleccionada en la grilla derecha, se muestran lo documentos que componen esta subfase con todas sus versiones. Desde aquí, si el usuario presiona el botón “Finalizar Proyecto”, el mismo solo quedará disponible a modo de consulta, es decir no se permitirá generar nuevas versiones de documentos ni modificar los actuales. Si presiona el botón “Abrir Documento”, previa seleccionando un documento en la grilla izquierda, o hace doble clic sobre el mismo, el sistema abrirá el documento Word que contiene la información. Estos documentos, tendrán en su versión original una breve introducción, la cual permitirá al usuario

comprender rápidamente que información volcar en el mismo. En la Figura 5 se muestra el contenido original del documento seleccionado en la pantalla 4.

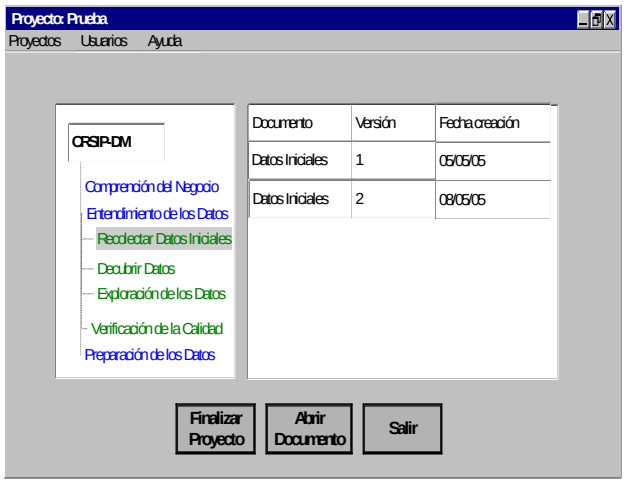

Figura 4: Pantalla de Administración de Proyectos

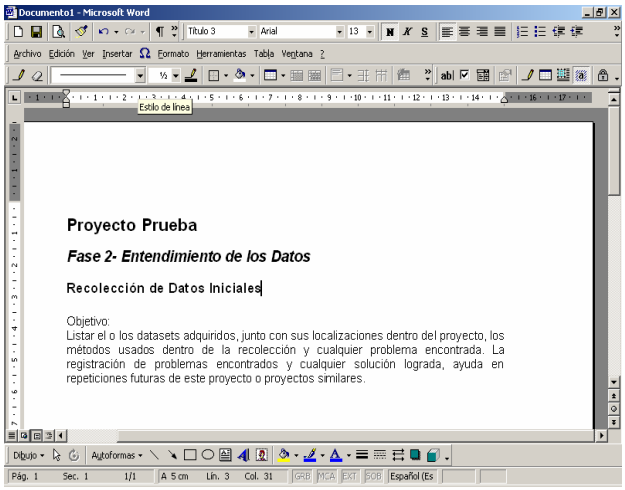

Figura 5: Contenido original del documento de Recolección de Datos Iniciales

La otra función a mostrar es la de asignación y consulta de usuarios asignados a un proyecto. Esta asignación se hace en base al perfil del usuario. Por ejemplo, un usuario con perfil de “Supervisor” puede ser asignado como responsable del proyecto completo, mientras que un usuario con perfil de “Desarrollador” solo puede estar asignado a una o mas Subfases del proyecto. La figura 6 muestra la pantalla de asignación de usuarios al proyecto. La misma posee dos grillas, en la grilla ubicada del lado izquierdo se pueden observar a todos los usuarios dados de alta en el sistema junto con su perfil, y en la ubicada del lado derecho se muestran todas las fases y subfases del proyecto junto con su responsable asignado. Para vincular a un usuario con algún elemento del proyecto, se debe seleccionar con doble clic un usuario de la grilla de la izquierda y luego se debe realizar una labor similar con los componentes del sistema descriptos en la grilla de la derecha (el orden de esta selección es indistinto). Ambos valores seleccionados se reflejan en los campos de texto descriptos debajo de cada grilla. Par vincular estos elemento, con los valores seleccionados, solo resta presionar el botón “Vincular” para que el sistema establezca la relación, lo cual queda reflejado en la columna usuario de la grilla derecha. Por último los botones de “Aceptar” y “Cancelar” permiten salir de esta pantalla, el primero guardando los cambios en la base de datos del sistema y el segundo perdiendo todos los vínculos establecidos en la sesión. Para poder observar con mayor claridad que proyectos tiene asignado un usuario, además de poder consultarlo desde la pantalla de asignación de usuarios, el sistema brindará un reporte donde se detalle esta información. En la figura 7, se muestra un reporte donde puede observarse los distintos usuarios asignados a los proyectos.

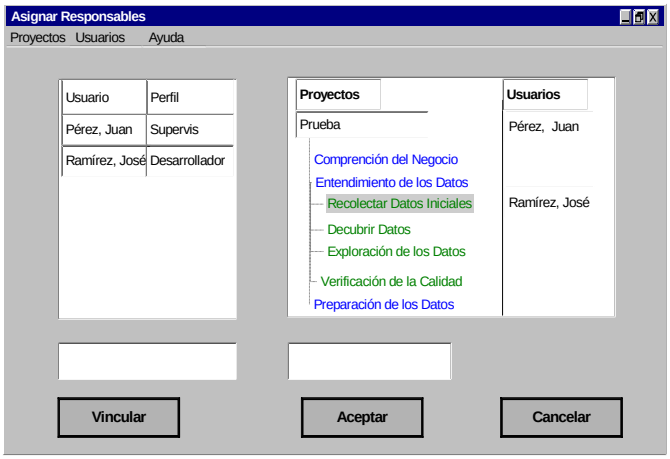

Figura 6: Pantalla de Asignación de Usuarios al Proyecto

Reporte de Proyectos Asignados

Fecha : 15/05/2005

Id. Usuario	Apellido	Nombre	Proyecto
1	Pérez	Juan	Prueba
2	Ramírez	José	Prueba

Figura 7: Reporte de Proyectos Asignados

6- Conclusión y Futuras Líneas de Investigación

El aportar una herramienta de gestión de documentos para la metodología CRISP-DM, que además proporciona un módulo de ayuda en línea, permite a quienes desarrollan el proyecto y están ya familiarizados con la metodología poder llevar a delante sus tareas de una forma mas aliviada. Por otra parte, para los desarrolladores novatos o junior el hecho de contar con una herramienta que tenga predefinidos todos los documentos a generar identificados en función de la fase de la metodología en que se encuentren ubicados y con la facilidad de contar con un módulo de ayuda en línea, que le aporte información sobre cual es el objetivo del documento y cuales son sus contenidos básicos del mismo, le permite una rápida entrada al equipo de desarrollo, haciendo que la curva de aprendizaje del mismo sea mucho mas suave. Además de los aportes indicados para los desarrolladores del proyecto, el contar con una herramienta que permita gestionar de forma centralizada los proyectos, constituye un medio de consulta fundamental para los niveles directivos de la organización. Quienes necesitan conocer el estado de cada uno de los proyectos que se están desarrollando. A este nivel se requiere poder ver que actividades o fases de la metodología se han completado y quienes o que desarrollador realizo la tarea, asimismo, el hecho de poder ver la cantidad de versiones generadas para un determinado documento dentro de un proyecto puede permitir sacar conclusiones respecto de su complejidad y tamaño. También se puede evaluar la habilidad de los distintos desarrolladores en función de la cantidad de versiones de documentos que generan y el tiempo que les lleva terminar su tarea.

Como líneas de investigación futura se han identificado: (a) la incorporación de nuevas metodologías a la herramienta software generada, esto podrá hacerse mediante la incorporación de un nuevo módulo que permita la parametrización de las tablas que contengan información sobre las fases, subfases y documentos a generar, (b) la ampliación de las funcionalidades del sistema respecto a la metodología CRISP-DM, aportando un módulo de sistema experto [García-Martínez y Britos, 2004] que permita, en función de las características del proyecto a tratar, determinar que fases y actividades deberían completarse. Facilitando de está forma la tarea de los desarrolladores (sobre todo a los novatos). Este punto de ampliación no será fácil extenderlo a otras metodologías debido a que para poder hacerlo se deberá contar con un conjunto de expertos dispuestos a brindar su experiencia para incorporarla al nuevo sistema y (c) la ampliación de la herramienta para que puede ser ejecutada desde múltiples plataformas hardware y software.

7- Referencias

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R. 1999. *CRISP-DM 1.0 Step-by-step data mining guide*. En www.crisp-dm.org. Página web vigente al 03-07-05.

CRISP, 1997. *CRoss Industry Standard Process for Data Mining Amsterdam SIG Workshop*. En http://www.crisp-dm.org/SIG/amsterdam.htm. Página web vigente al 03-07-05.

Diez, E., Britos, P., Rossi, B. y García-Martínez, R. 2003. *Generación Asistida del Mapa de Actividades de Proyectos de Desarrollo de Software*. Reportes Técnicos en Ingeniería del Software. (5)1:13-18.

García Martínez, R. y Britos, P. 2004. *Ingeniería de Sistemas Expertos*. Editorial Nueva Librería. Buenos Aires.

Gesmet, 2001. *Gestor metodológico de la metodología Métrica Versión III* (versión 1.0.0 del 13 de marzo de 2001). Desarrollado por Getronics. En www.getronic.com. Página web vigente al 03-07-05.

IBM, 2005. *Rational Unified Process*. En http://www-306.ibm.com/software/awdtools/rup/. Página web vigente al 03-07-05.

Métrica, 2004. *Métrica III: Metodología de Planificación, Desarrollo y Mantenimiento de sistemas de información*. Consejo Superior de Informática y para el Impulso de la Administración Electrónica Española. En http://edic.lsi.uniovi.es/metricav3/. Página web vigente al 03-07-05.

Pressman, R. 2002. *Ingenieria del Software: Un Enfoque Práctico*. 5ta edición. McGraw-Hill. Madrid.

Somerville, I. 1998. *Ingeniería del Software*. Addison Wesley Iberoamericana. Madrid.

ERROR: syntaxerror
OFFENDING COMMAND: --nostringval--

STACK:

(
CACIC-2005-Gestion-Asistida-de-Documentos-en-Metodología-de-Explotacion-de-Información-1
)
/Title
()
/Subject
(D:20050926143103)
/ModDate
()
/Keywords
(PDFCreator Version 0.8.0)
/Creator
(D:20050926143103)
/CreationDate
(pbritos)
/Author
-mark-